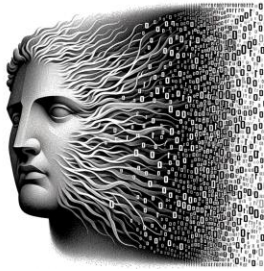
ASPLOS 2024



Atalanta:

A Bit is Worth a “Thousand” Tensor Values

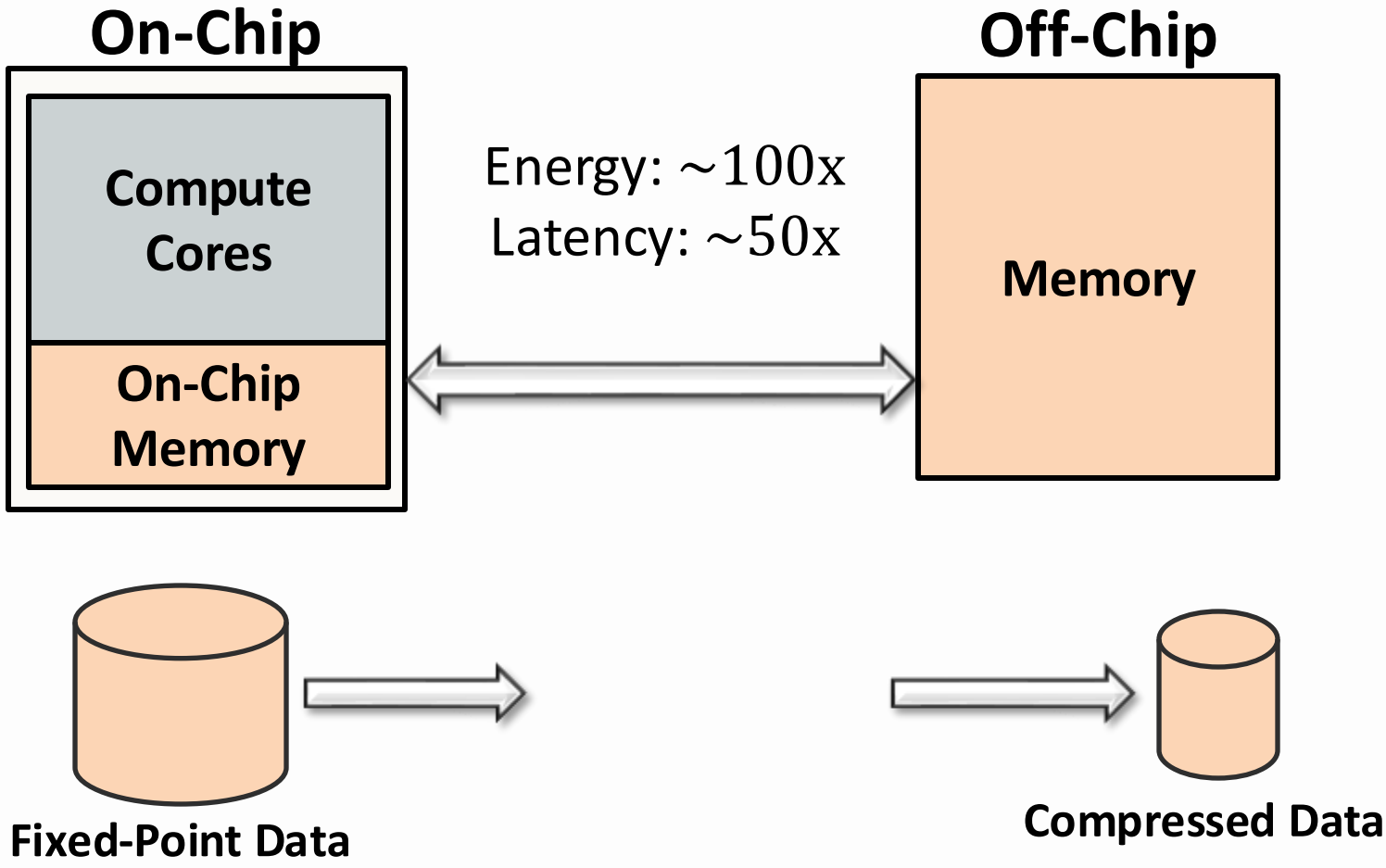
Alberto Delmas Lascorz, Mostafa Mahmoud, Ali Hadi Zadeh, Miloš Nikolić,
Kareem Ibrahim, **Christina Giannoula**, Ameer Abdelhadi,
and Andreas Moshovos[1,2]

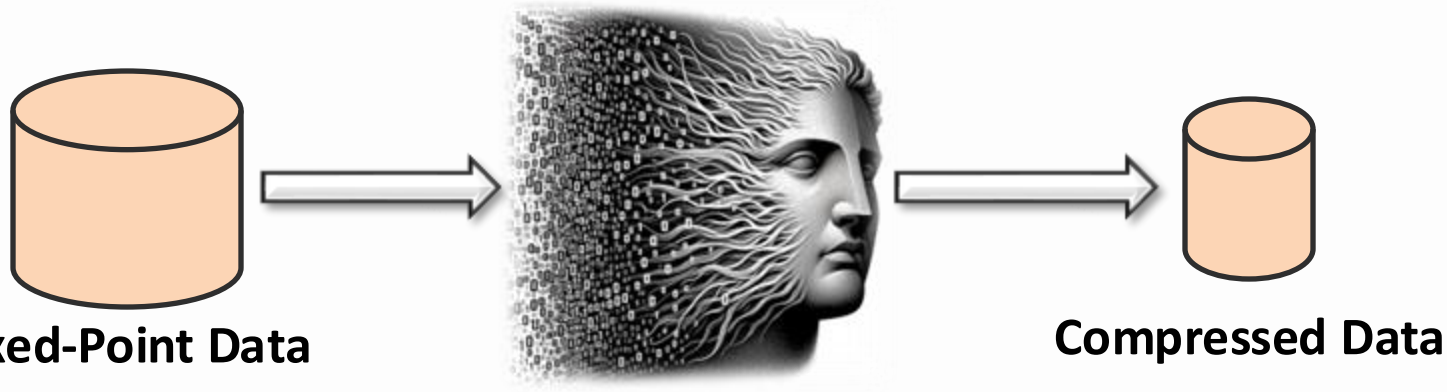
UNIVERSITY OF TORONTO

VECTOR INSTITUTE

McMaster University

1QBit


On-Chip
Compute Cores
On-Chip Memory
Energy: ~100x
Latency: ~50x
Off-Chip
Memory
Fixed-Point Data
Compressed Data



Inference & Training

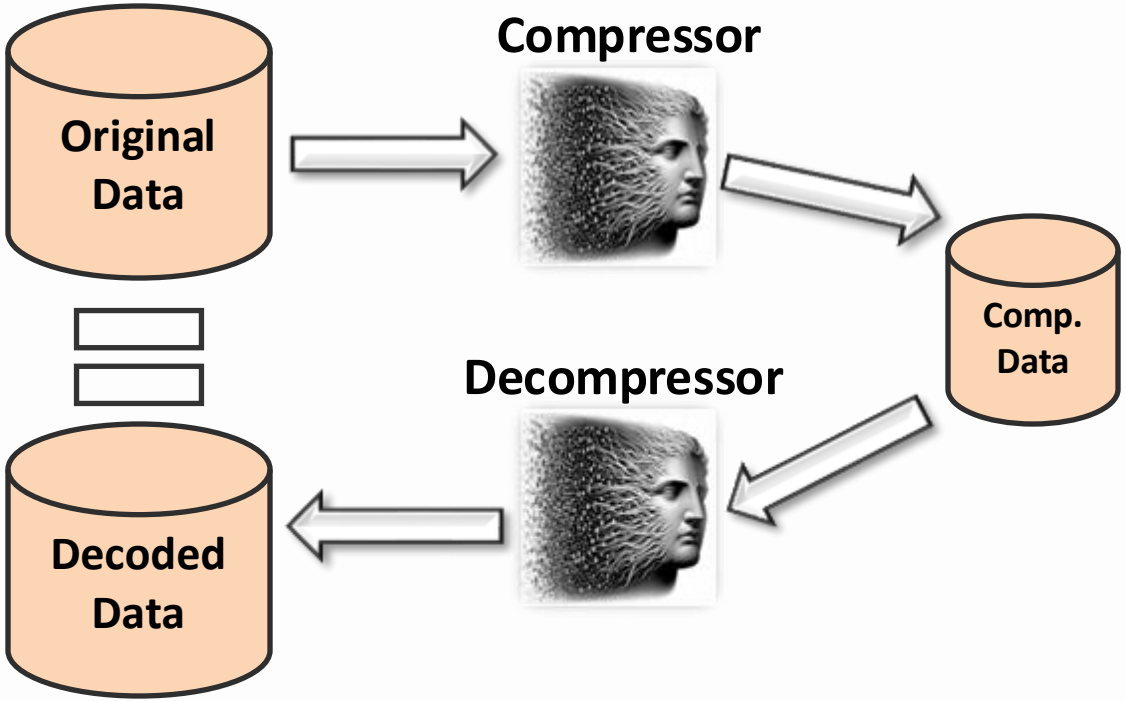

Transformers


Classification


Recommendation

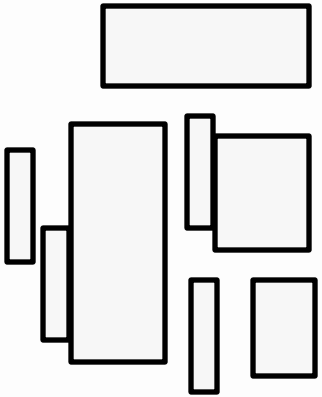

Detection

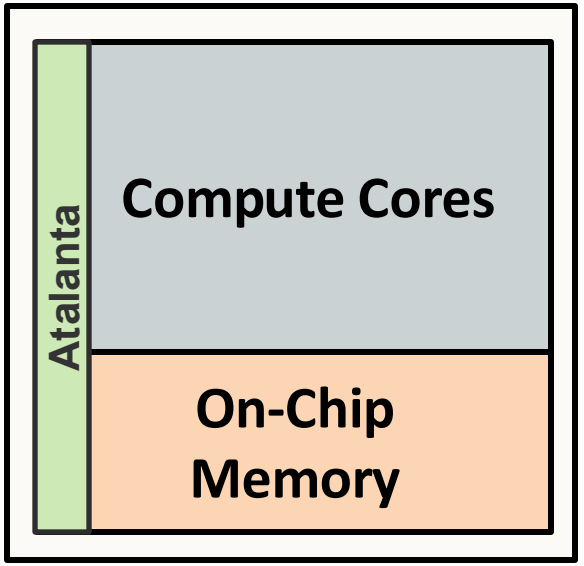
Lossless


Compressor
Original Data
Comp. Data
Decompressor
Decoded Data

Lossless

Transparent

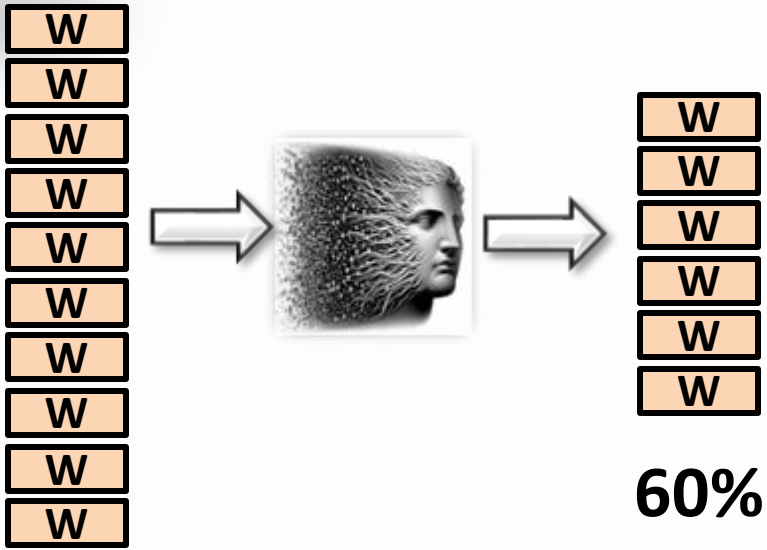




Lossless

Transparent

High Compression

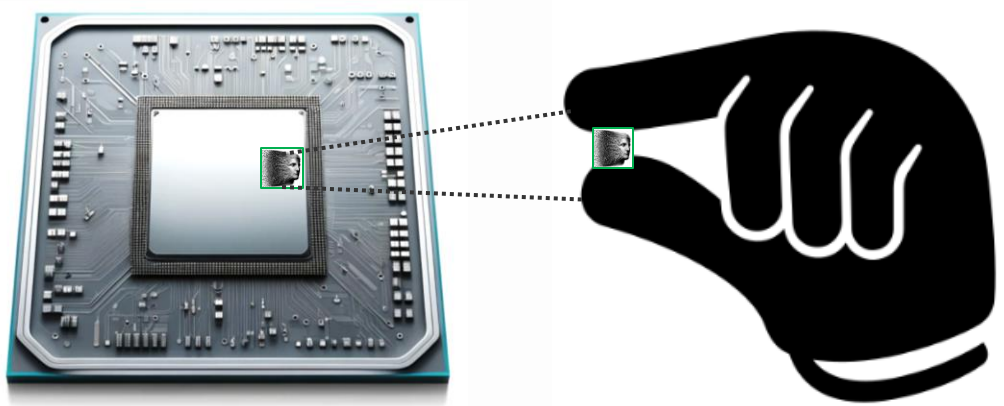


Lossless

Transparent

High Compression

Low Cost



Roadmap

0 Motivation

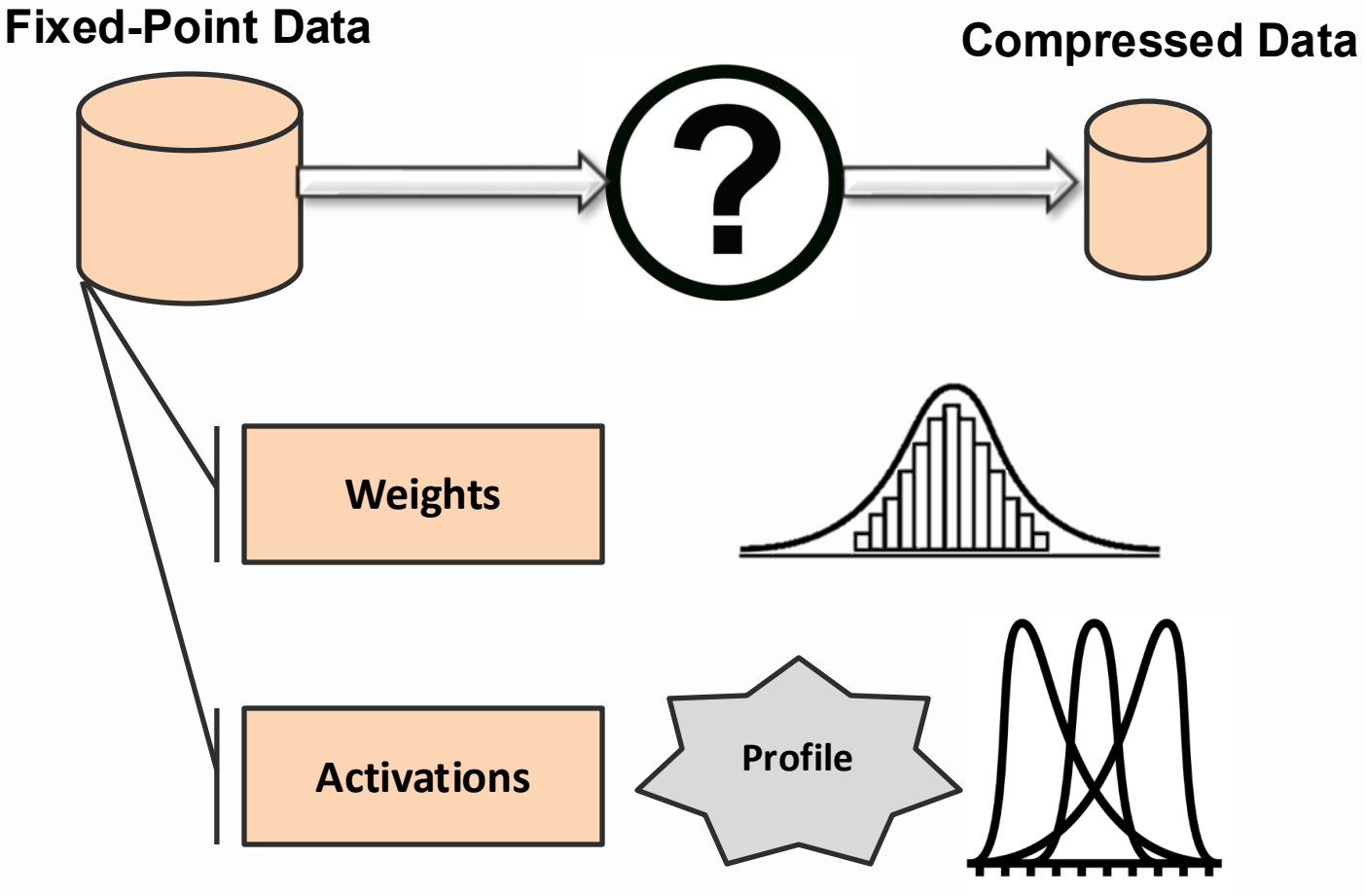
1 Key Idea

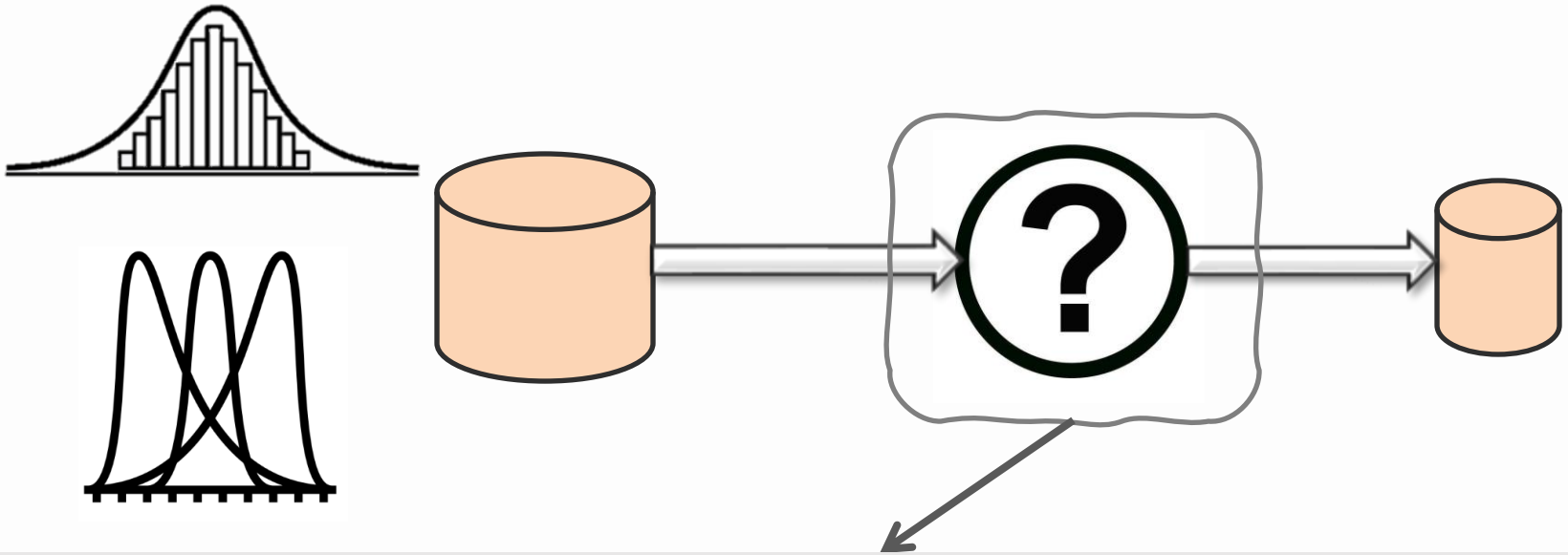
2 Atalanta's Software

3 Atalanta's Hardware

4 Results

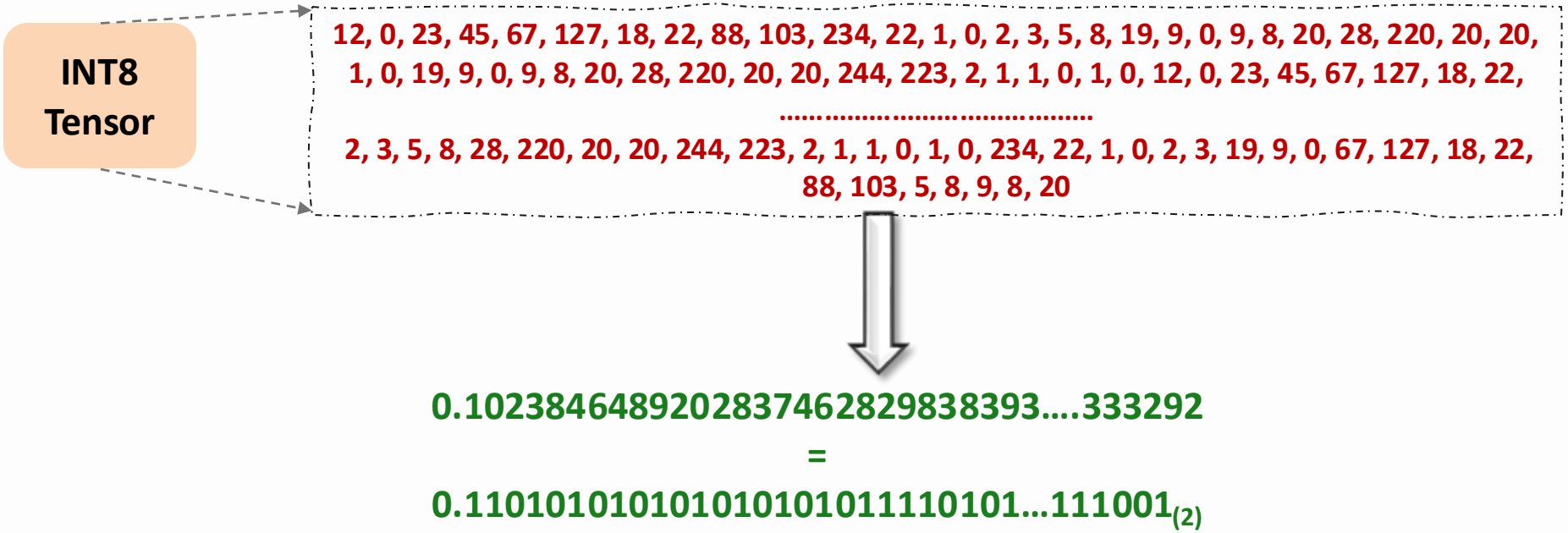
5 Conclusion


Fixed-Point Data
Compressed Data
?
Weights
Activations
Profile

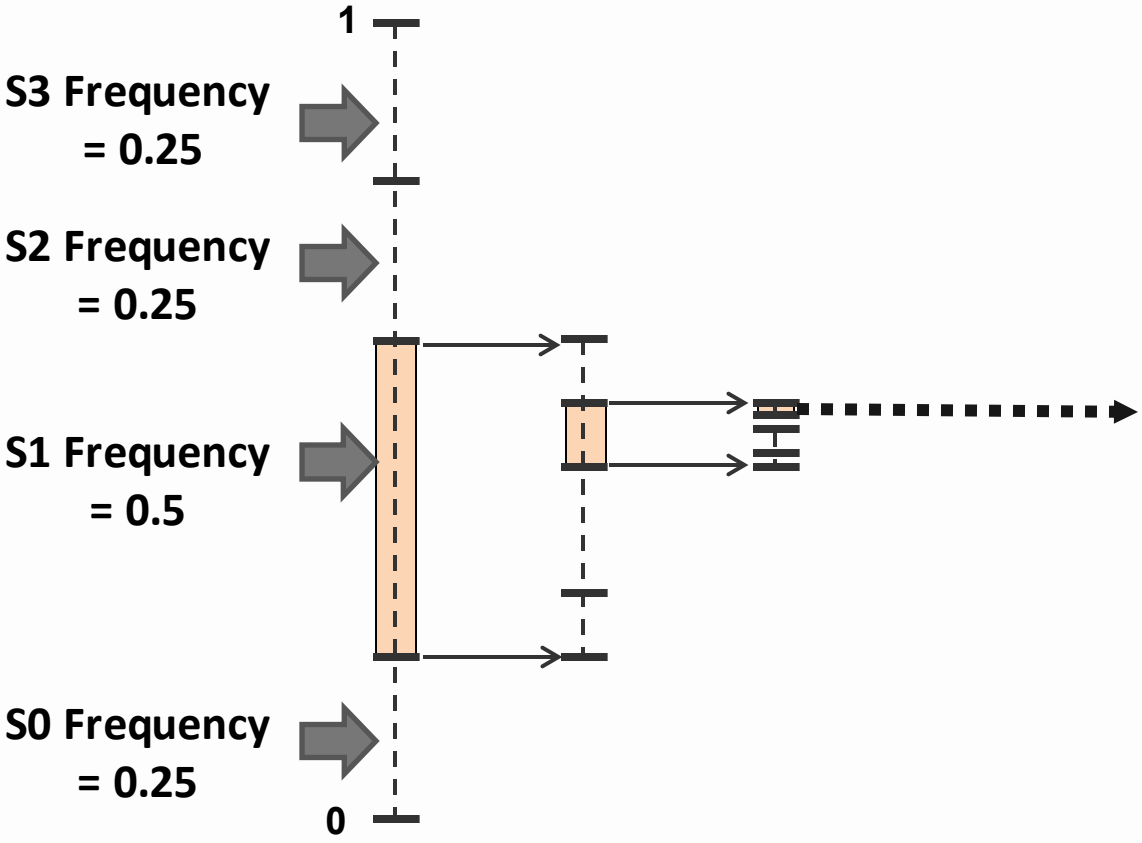


Arithmetic Coding?

Arithmetic Coding



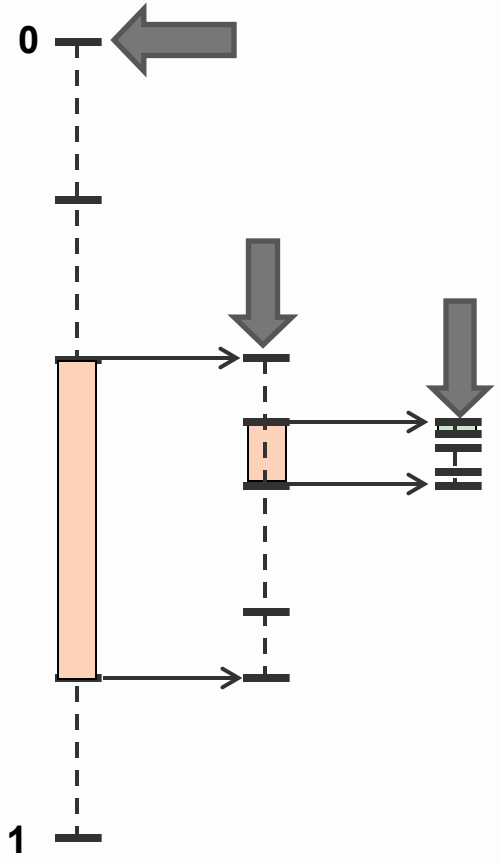
Arithmetic Coding – INT2 Example



$0.11001010101\ldots101_{(2)}$

NOTHING IS PERFECT

Arithmetic Coding – Probability Table



Symbol	Frequency
S_1	P_1
S_2	P_2
.	.
.	.
.	.
.	.
.	.
.	.
S_N	P_N

INT8

256 Entries

93.6% DRAM Power

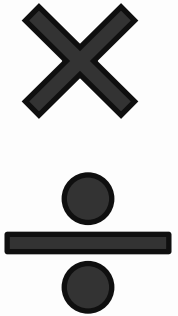
Arithmetic Coding – Precision

Expandable Precision

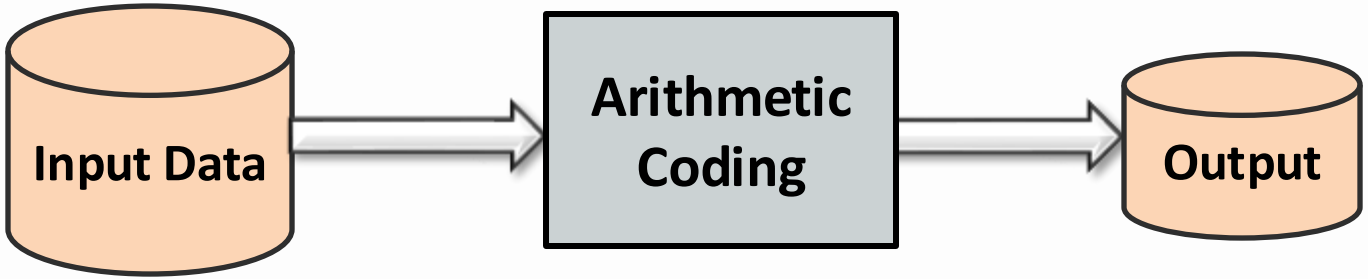
0.110010…….101$_{(2)}$

=

Costly Operations



Arithmetic Coding – Bandwidth



1 bit/cycle

Atalanta – Key Idea

VALUE = BASE + OFFSET

int8 3 bits 5 bits

01010111 = 01000000 + 10111

Apply AC

Store

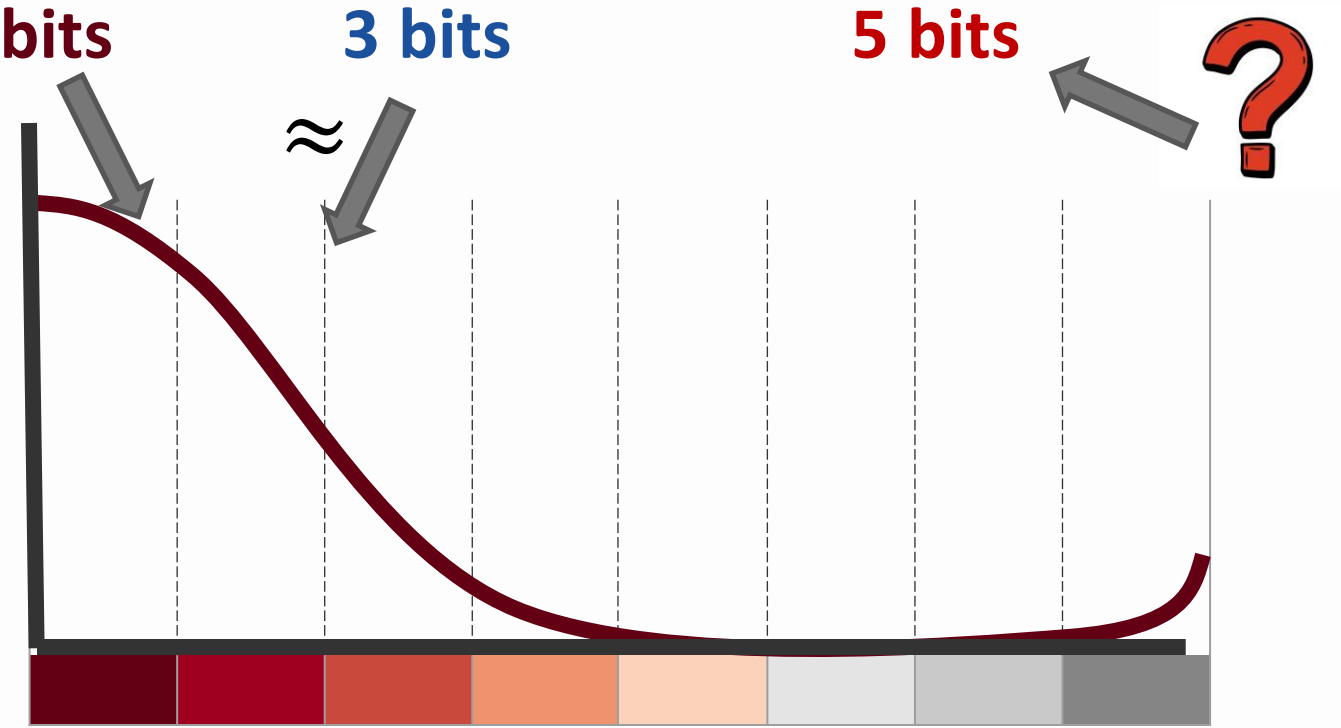
256 Entries

8 Entries

Atalanta – Key Idea

VALUE = BASE + OFFSET

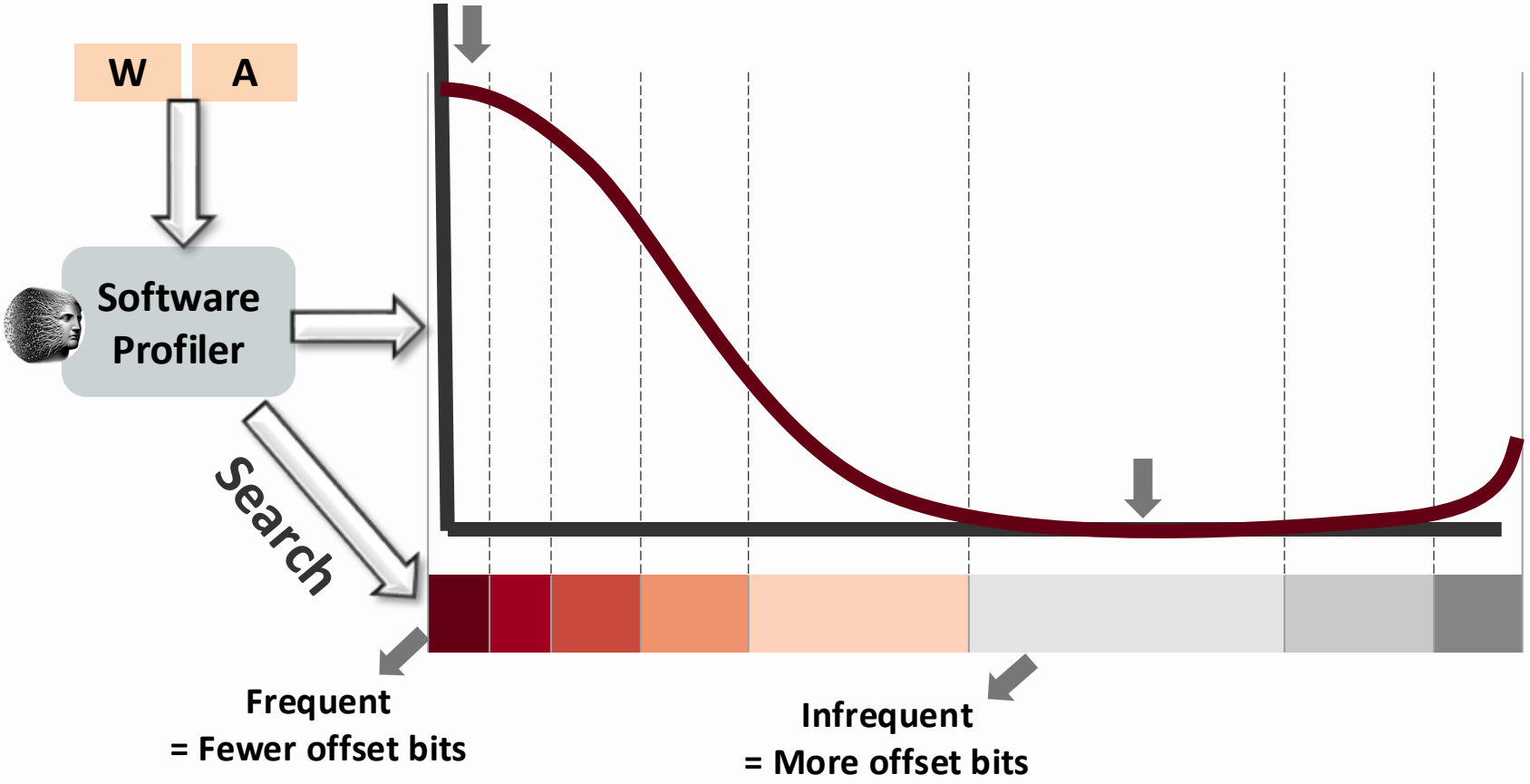
8 bits **3 bits** **5 bits**



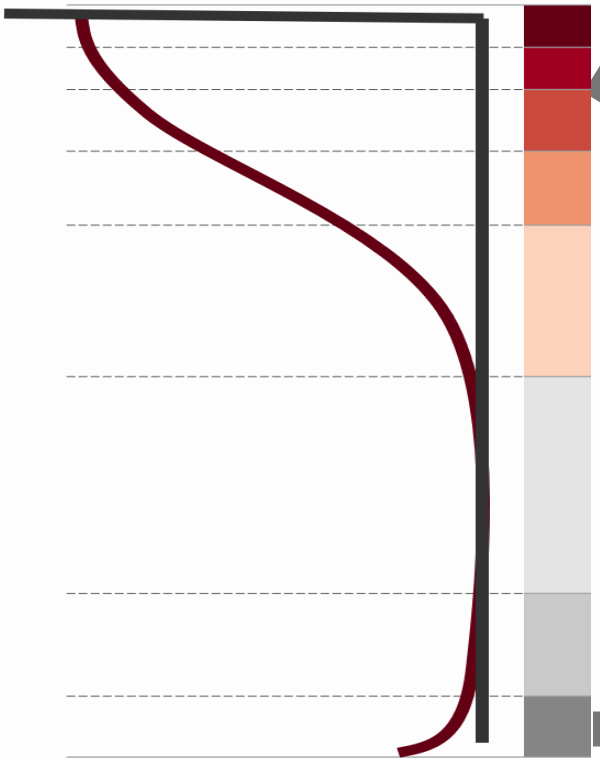
Software
Profiler

Hardware Runtime
[Encoder/Decoder]

Atalanta – Software Profiler



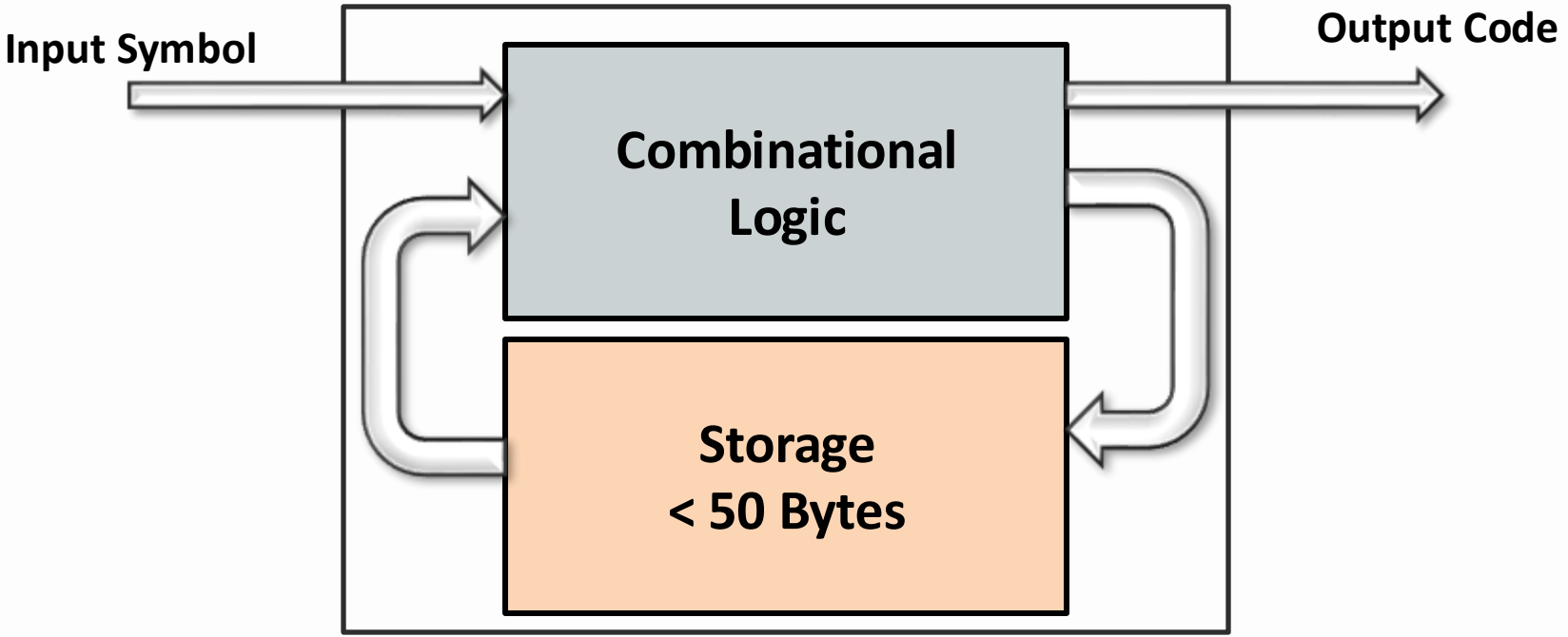
Atalanta in Software



8b	3b	10b
Min Value	#Offset Bits	Probability Limit
V_{min_1}	O_1	t_{high_1}
V_{min_2}	O_2	t_{high_2}
. . . .	. . . .	. . . .
V_{min_N}	P_N	t_{high_N}

8/16 Entries

Atalanta Encoder in Hardware



Atalanta During Inference Evaluation

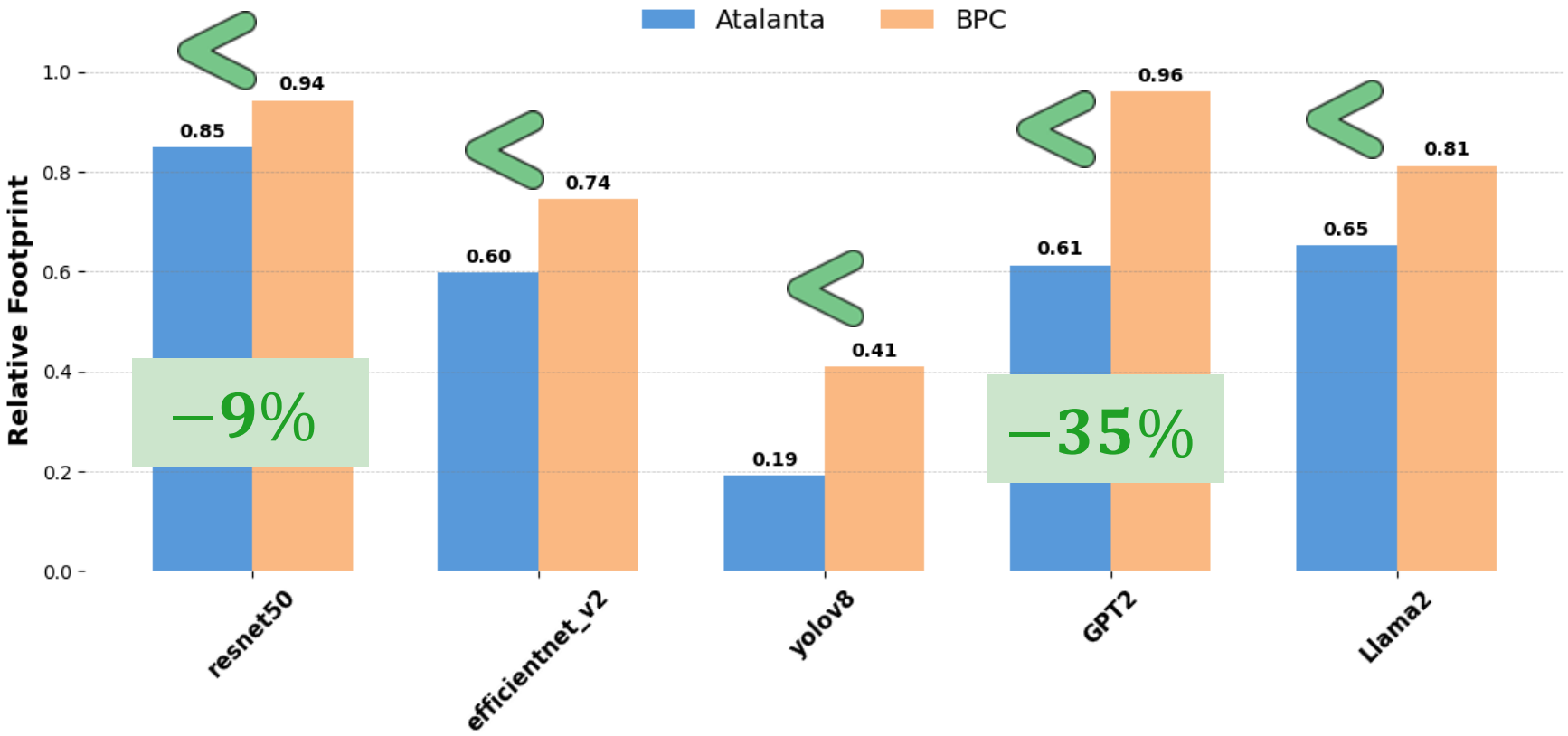
VS

EFFICIENTNET v2

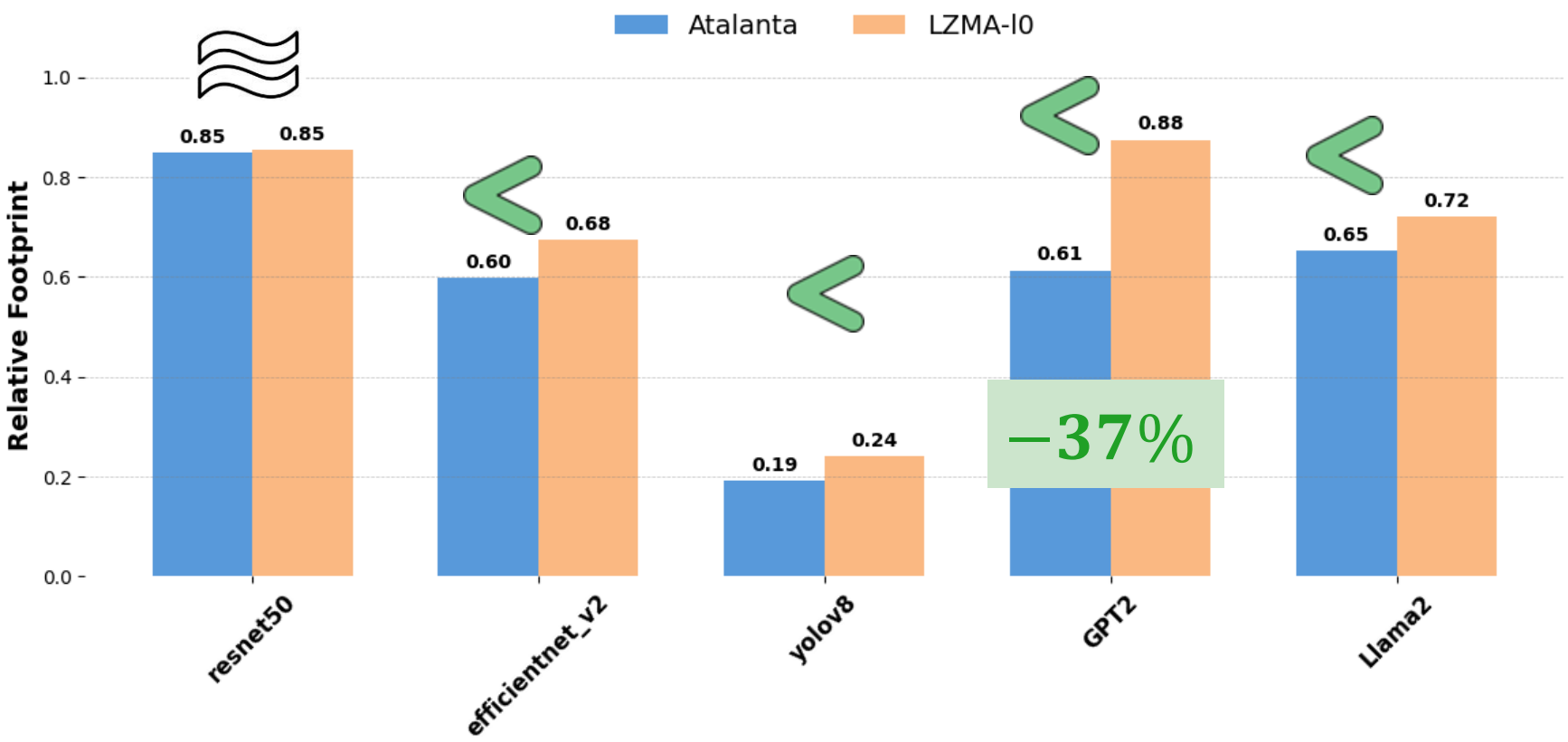
Bit Plane Compression (BPC)

YOLOV8

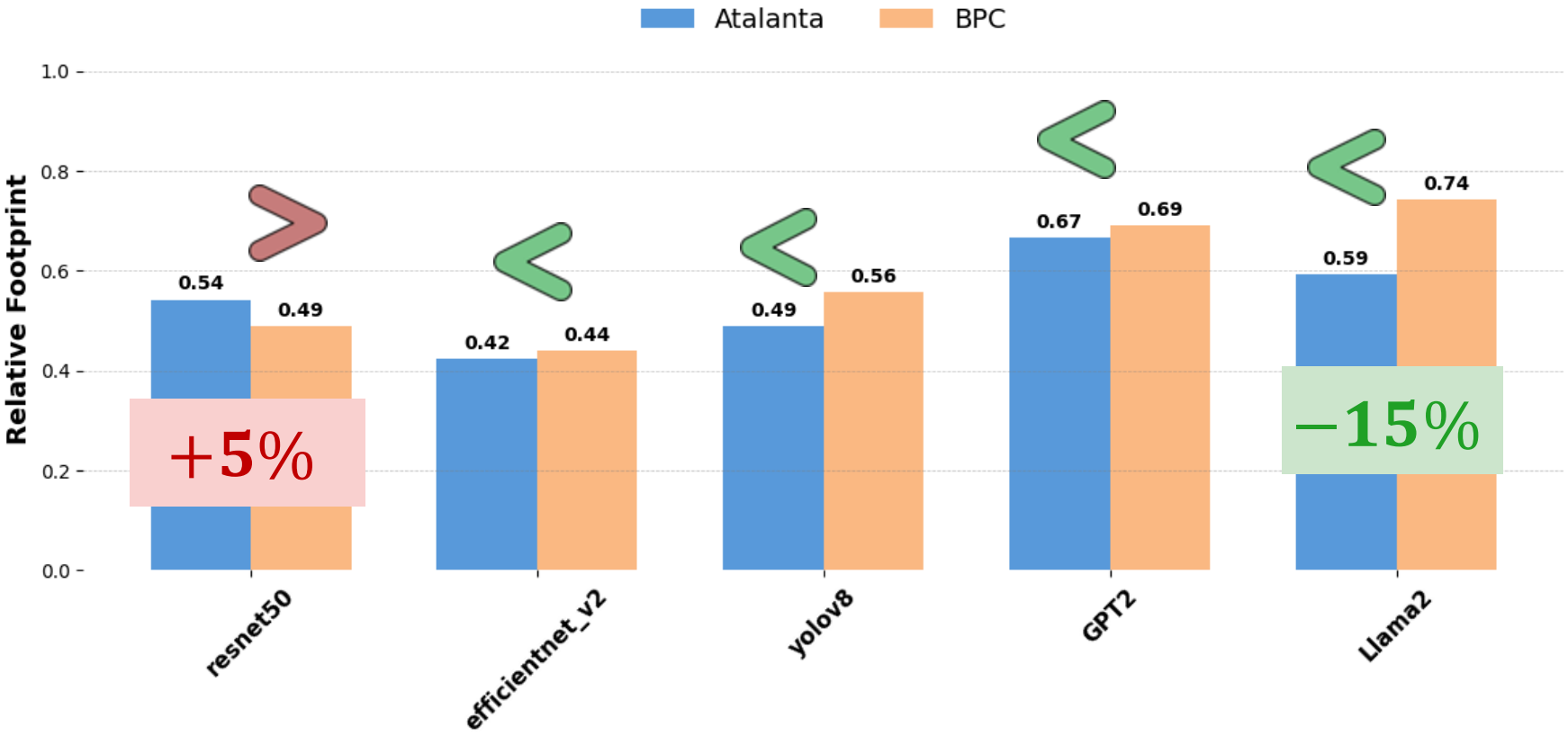
Compression Performance: Weights



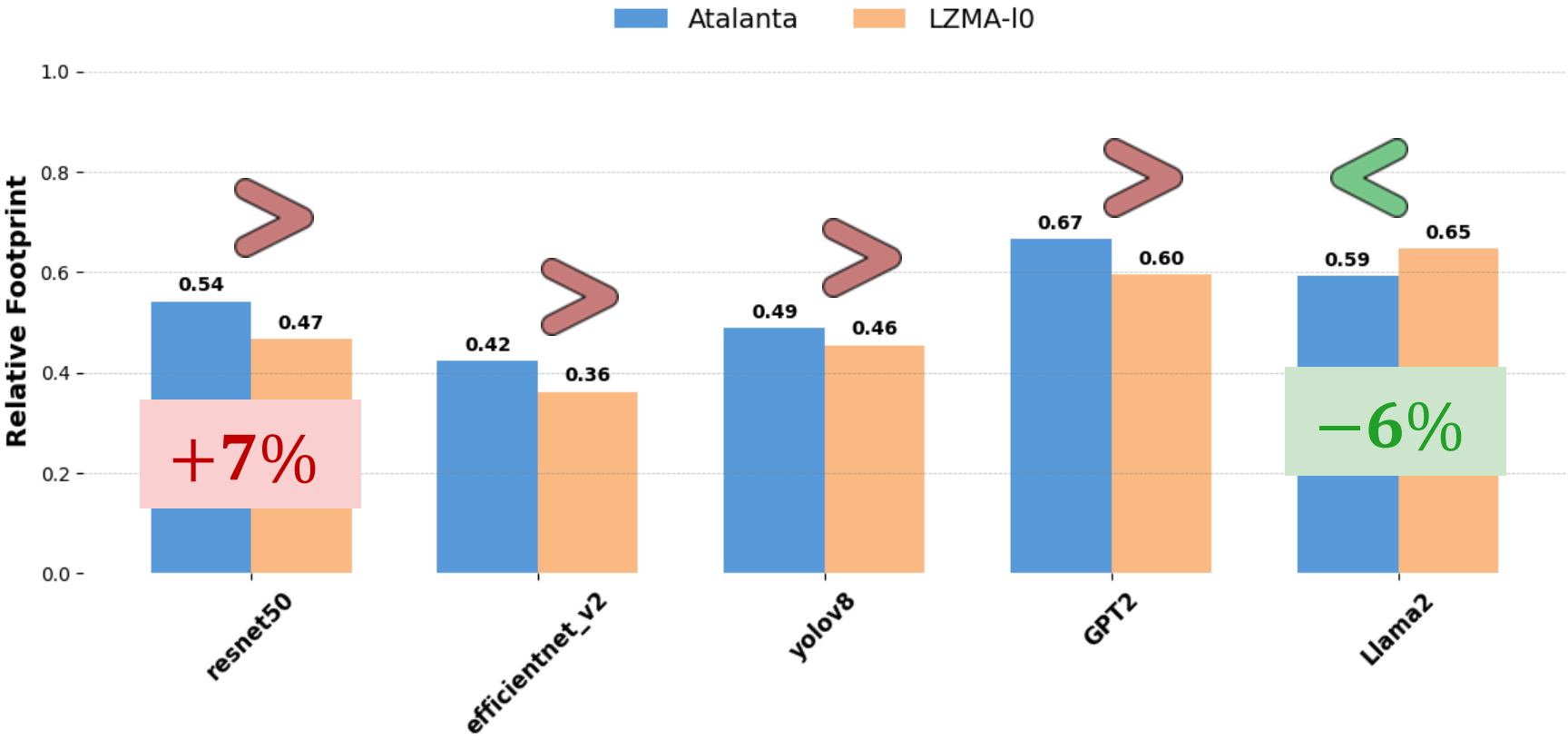
Compression Performance: Weights



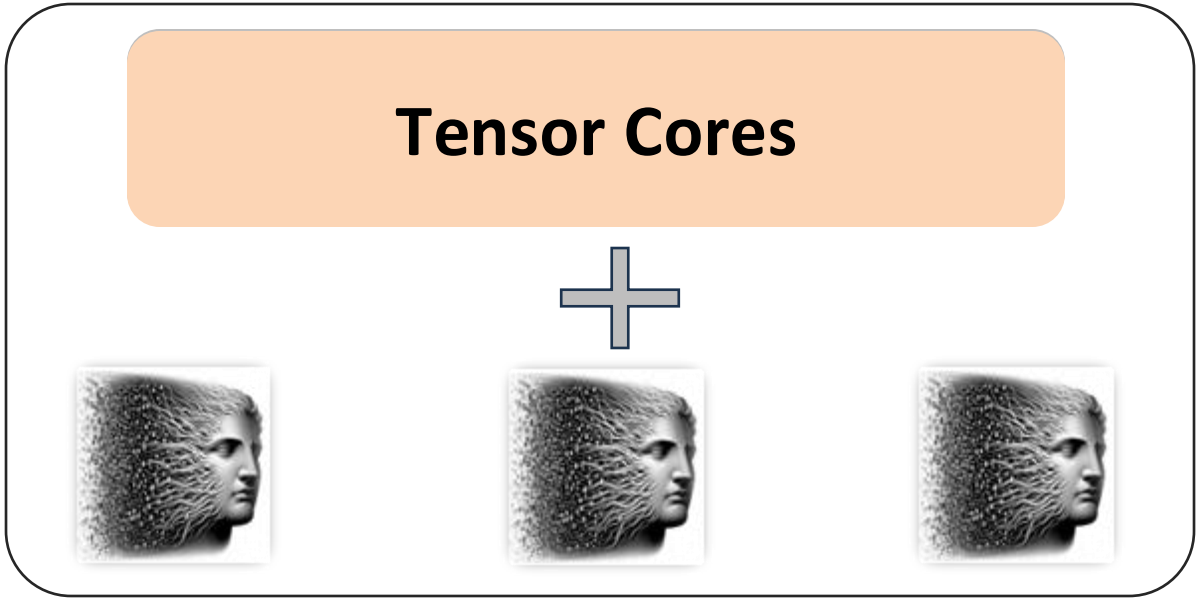
Compression Performance: Activations



Compression Performance: Activations



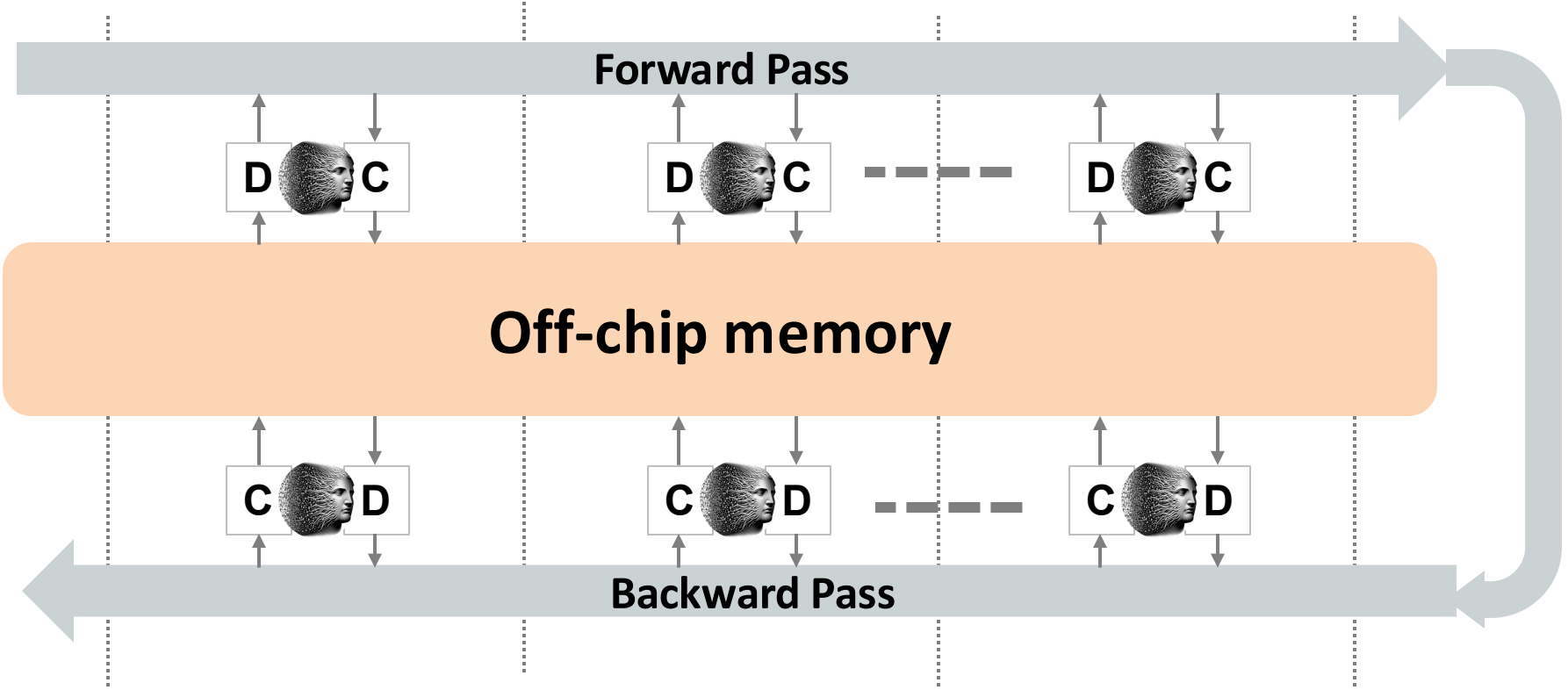
Energy Efficiency and Speedup



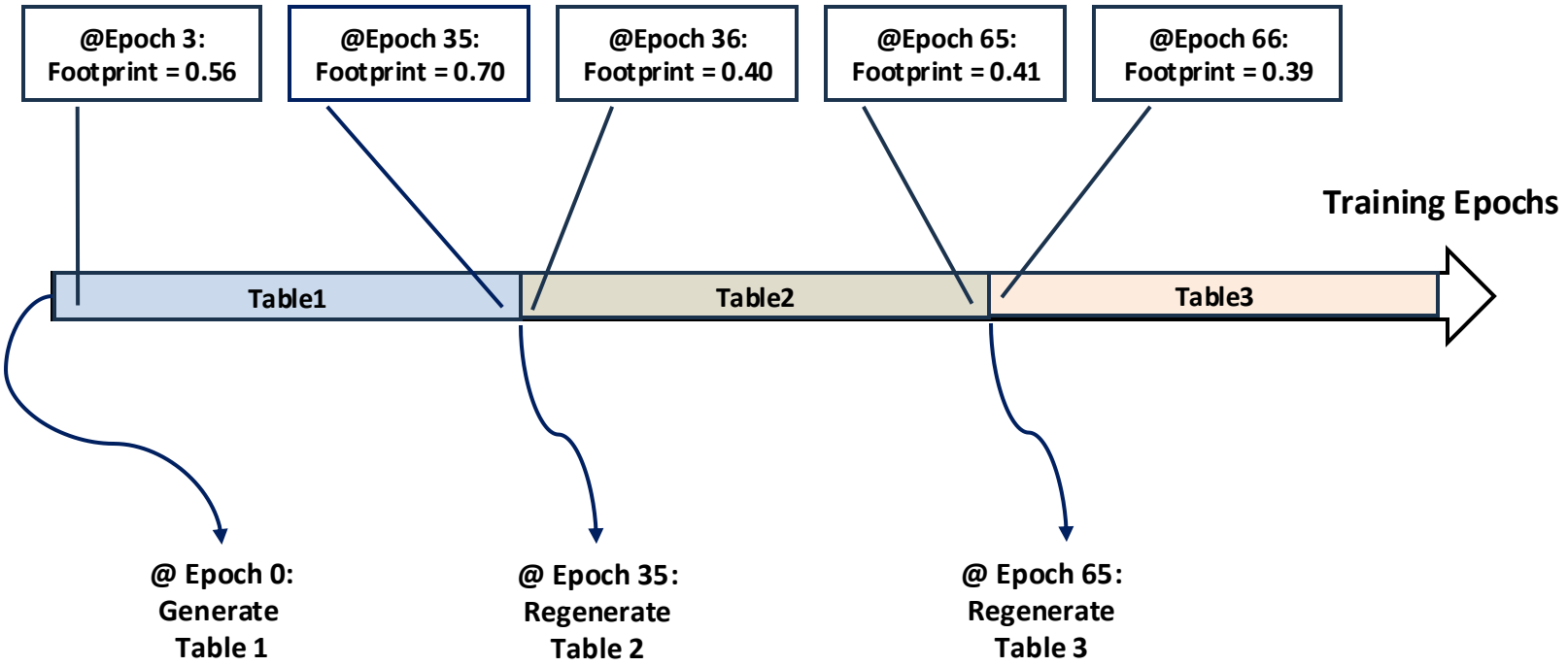
Speedup: 1.44x

Energy Efficiency: 1.37x

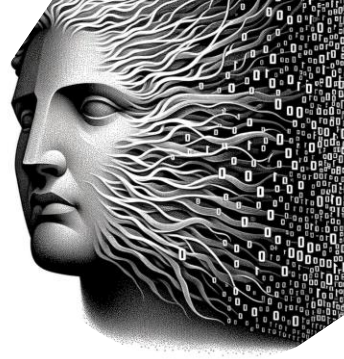
Training for Fixed-Point Inference

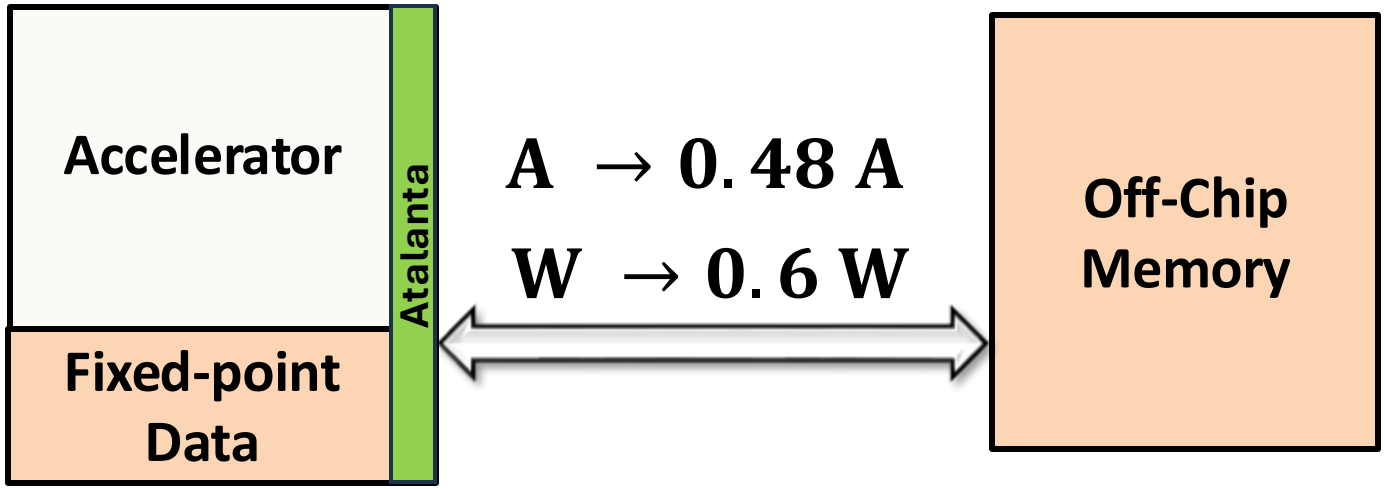


Compression Performance During Training

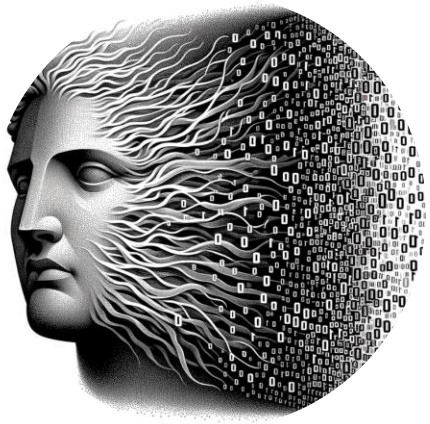


Conclusions





Lossless | **Transparent** | **Low-cost**



Thank you!